

NYELVÉSZETI PROBLÉMÁK A TULAJDONNÉV-FELISMERÉS TERÜLETÉN*

SIMON ESZTER

1. Mi a tulajdonnév-felismerés?

A tulajdonnév-felismerés a számítógépes nyelvészet egyik területének, az információkinyerésnek egy alfeladata. De mivel foglalkozik a számítógépes nyelvészet, és mi a célja az információkinyerésnek?

A *számítógépes nyelvészet* (computational linguistics) a kognitív tudományok közé tartozik, átfedésben van a mesterségesintelligencia-kutatással, melynek elsődleges célja az emberi kogníció számítógépes modellálása. A számítógépes nyelvészeti kutatások a nyelv szerkezetének gépi modellezésére irányulnak, céljuk a természetes nyelvek számítógépes feldolgozása.

Az *információkinyerés* (information extraction) a számítógépes nyelvészet egyik fontos és mostanában meglehetősen felkapott alterülete. Célja, hogy a számítógép által olvasható, ámde strukturálatlan szövegből gépi eszközökkel, automatikusan információt nyerjünk ki. Egy információkinyerő rendszer feladata, hogy automatikusan adatbázisba rendezze az adatokat, amelyek így már használhatók az adatok analizálására, összegzést kaphatunk belőlük természetes nyelvi jelenségekről, vagy bármilyen online eszköz bemenetül szolgálhatnak. A feladatok igen tág köre tartozik ez alá: például megtalálni az összes cégnevet egy szövegben, vagy kideríteni a szövegből, hogy ki ölt meg kicsodát, vagy egyáltalán milyen esemény történt, milyen szereplőkkel. A lényeg, hogy hatalmas mennyiségű szöveg átbogarászása helyett csak a számunkra fontos, specifikált információt kapjuk meg.

A *tulajdonnév-felismerés* (named entity recognition) az információkinyerés egy alfeladata. A cél a szövegben található olyan elemek megkeresése, amelyek a világ valamely entitására egyedi módon (unikusan) referálnak. Ezek a példák:

* A cikk megírásában nagy segítséget nyújtott Kornai András, Kenesei István és természetesen az ismeretlen lektor. Ezúton is köszönöm.

- (1) *Eötvös József Gimnázium*
- (2) *Eötvös*
- (3) *E5vös*
- (4) *EJG*
- (5) *a suli az utcában*

mind egy intézményre referálnak ugyan, de csak az első négy esetben használtunk egyedi jelölőt a megnevezésükre. A tulajdonnév-felismerés körébe a teljes tulajdonnevek, ezek különböző rövidítései, a becenevek és a mozaikszavak felismerése tartozik, de a szimpla köznévi frázisoké nem.

2. A nevek azonosítása és osztályozása

Egy szöveg nyelvi elemzése általában azzal kezdődik, hogy a szöveg szavait főnévként, melléknévként, igeként stb. azonosítjuk szótárak segítségével. Viszont a legtöbb szöveg tartalmaz neveket, amelyeket nem tud értelmes nyelvi egységként azonosítani a rendszer. Így tehát a tulajdonnév-felismerés nélkülözhetetlen lépése bármilyen szöveg nyelvi elemzésének, az eseménykivonatolásnak (event extraction) és a gépi fordításnak. Az, hogy a gépi fordító rendszer bizonyos szavakat vagy szósorokat nem tud névként azonosítani, sok fordítási hibának a forrása. Tipikus hiba például, amikor egy többrészes név részeit külön-külön fordítja le a gépi fordító. Olyan ez, mint amikor egy idiomatikus kifejezést szóról szóra akarnánk lefordítani: elveszteni az értelmét.

A különböző típusú szövegeket különböző kategóriájú nevek dominálják. Például biológiai témájú cikkekben jellemzően sok gén- vagy proteinnév, kémiaiakban pedig sok vegyületnév fog előfordulni. Egy általános újságcikkben viszont nagyrészt személyek, helyek és szervezetek nevei fognak nagy számban szerepelni. Ezek azok a névosztályok, melyekkel hagyományosan a legtöbb névklaszifikáló dolgozik.

A tulajdonnév-felismerés tehát két fő lépésből áll: először lokalizálni kell a szövegben a nevet, aztán besorolni egy előre definiált névosztályba. Tipikus névosztályok: a személy-, a hely-, az intézménynevek, a dátumok és egyéb időre referáló kifejezések, valamint a különböző mennyiségeket jelölő elemek. Például:

- (6) *Kosztolányi Dezső*
- (7) *1997. ápr. 5-én*

(8) *United Nations Educational, Scientific and Cultural Organization*

(9) *Déli-Shetland-szk.*

(10) *IBM*

(11) *Kiss János altábornagy utca*

(12) *Műgyetem*

(13) *500\$*

(14) *Kovács Pistike*

Ezekből a példákból jól látszik, hogy az angol *named entity recognition* terminus a szövegelemek nagyobb csoportját fedi le, mint a magyar tulajdonnév-felismerés, mégis ezt a kifejezést használjuk, mivel még mindig ez a legjobb közelítése a fordításnak. (A tulajdonnév-felismerés témája hivatalosan 1995-ben, a hatodik Message Understanding Conference-en bukkant fel először. A hetedik MUC-ra összeállított útmutató (Chinchor és mtsai 1998) az, amelyet a mai napig a legtöbben használnak vagy hivatkoznak a tulajdonnév-felismerés terén, melyben a felismerendő szövegrészek között a hagyományosan tulajdonnévnek tartott elemek mellett pénzügyi és százalékos kifejezések is szerepelnek.)

3. Korpuszépítés

A tulajdonnév-felismerő rendszerek legújabb generációjába a statisztikai modellek tartoznak, melyek a gépi tanulás módszereivel tanulják ki a szövegből az egyes tulajdonnevek jellemző tulajdonságait. Ehhez nagy mennyiségű szövegre, úgynevezett korpuszokra van szükség. A korpusz természetes nyelvű szövegek nagy és elvszerűen rendezett gyűjteménye, amely empirikus elemzésre ad lehetőséget. Egy manuálisan annotált korpusz, vagyis egy olyan szöveghalmaz, amelyben meghatározott szabályok alapján kézzel megjelölték a neveket, és kategóriájuknak megfelelően felcímkezték őket, jól használható gépi tanuláson alapuló névklasszifikáló rendszerek tanítására és szabványos kiértékelő korpuszként. Egy gépi tanuló algoritmus egy ilyen korpuszból tanulja meg a paramétereit automatikus módon, és szokásosan egy algoritmus kiértékelése is ilyen korpussszal való összevetés útján történik. Nagyon fontos a szövegek jó megválasztása, mert nagy hatással van a gépi tanulás eredményére. Ha olyan statisztikai modellt szeretnénk a szöveg alapján építeni, amely általános és specifikus szövegen is jól megállja a helyét, akkor a korpusz jellegét úgy célszerű meghatározni, hogy a korpuszt

alkotó szövegek témájukat tekintve heterogének legyenek, és az egész korpusz és egyes részei önállóan is kellően nagyok legyenek.

Tulajdonnév-címkéző rendszerek fejlesztéséhez tehát elengedhetetlenül szükséges egy kellően nagy méretű, tematikusan heterogén, konzisztens annotálási szabályzaton alapuló, manuálisan feljelölt korpusz. Ennek létrehozására indult a HunNER projekt három konzorciumi tag: a BME Média Oktató és Kutató Központ (MOKK), a Szegedi Tudományegyetem Informatika Tanszékcsoportja és az MTA Nyelvtudományi Intézete részvételével (Simon és mtsai 2006). A projekt során kialakítottunk egy egységes annotációs útmutatót, amelyben a nemzetközi szinten használt útmutatókat közös munkával konzisztens rendszerré ötvöztük, és a magyar nyelvre hangoltuk. A tulajdonnevek speciális tulajdonságai miatt a munka során jó pár nyelvészeti problémával kellett szembenéznünk; ezekről lesz szó a dolgozat további részében.

4. A tulajdonnév definíciója

A legelső probléma az, hogy mely szövegelemeket tekintjük annotálandónak. Kikötöttük, hogy csak tulajdonneveket annotálunk – de hogyan definiáljuk a tulajdonneveket? Tulajdonnévnek nevezzük azt a kifejezést, ami unikusan, vagyis egyedi módon referál a világ valamely entitására. Tehát, ahogy ezt már korábban említettem, nem annotálunk olyan frázisokat, amelyek ugyan a világnak valamely egyedi részére utalnak, de nem egyedi módon. A tulajdonnév-felismerés feladatai körébe nem tartozik bele a köznévi frázisok jelölése, még akkor sem, ha azok a világnak egy egyedi entitására referálnak; ahogy a nyelvészetébe sem, hogy a jelöletek egyediségéről bármit is állítson.

Miről ismerjük fel a tulajdonneveket? Az iskolás válasz: a nagy kezdőbetűről és a névelő hiányáról. De mi a helyzet azokkal a helyesírási rendszerekkel, amelyekben minden főnevet nagybetűvel írnak, vagy azokkal az írásrendszerekkel, amelyekben nincs nagybetű, illetve azokkal a nyelvekkel, amelyekben nincs névelő? Nyilvánvalóan ezekben is felismerhetők a nevek. A fenti két tulajdonság tehát fontos, de nem egyedüli indikátora a tulajdonneveknek az írott szövegben. A magyarban is vannak esetek, amikor szerepel határozott névelő a tulajdonnév előtt, vagy amikor kisbetűvel írjuk a tulajdonnevet vagy legalább egy részét, például az utcanevek köznévi elemét.

A kisbetű-nagybetű kérdés felvet egy másik fontos, állandóan visszatérő problémát: a tulajdonnév-felismerés és a helyesírás viszonyát. A legeggy-

szerűbb megoldás erre az, ha valaki csak a helyesírási szabályzatnak megfelelően írt neveket jelöli, vagyis a következő két példa közül csak a másodikat:

(15) *Marczibányi téri általános iskola*

(16) *Marczibányi téri Általános Iskola*

Ezzel a megközelítéssel szemben a HunNER korpusz annotálása során azt tartottuk szem előtt, hogy olyan névklasszifikációs útmutatót hozzunk létre, amely nem függ az akadémiai helyesírási szabályzat éppen aktuális irányadásaitól.

A kis- és nagybetűs használat ingadozásának oka ebben a példában az, hogy kevésbé tulajdonnévszerű, mert a tulajdonnév „megkülönböztető eleme” (J. Soltész 1979) csak a működési hely feltüntetése. Hasonlóan nehéz eldönteni, hogy tulajdonnevek-e vagy sem azok az esetek, amelyekben a megkülönböztető elem csak egy szám. Például tulajdonnevek-e a következő példák?

(17) *55-ös szavazókör*

(18) *70. sz. Postahivatal*

A (15)-(16) típusú példák esetében az is problémát okoz, hogy meddig tart a tulajdonnév, vagyis hogy a köznévi elemek is hozzátartoznak-e. A nevek jó része tartalmaz közzsói tagot/tagokat is, amelyek a név határán helyezkednek el; nehéz megmondani, hogy egészen pontosan melyek részei a névnek, és melyek nem. Elsősorban a hely- és a szervezetneveknél jellemző, hogy ingadozó státuszú közneveket találunk előttük-utánuk. Mivel a célunk egy általános annotálási útmutató megírása volt, nem hozhattunk egyedi döntést minden esetre, így azt a szabályt mondtuk ki, hogy a közvetlenül a tulajdonnév előtt vagy után álló, magyarázó, deskriptív funkciójú köznévi tagok a névvel együtt annotálandók. Minél több információt hordoz, vagyis minél inkább pontosít a köznévi tag, annál szorosabban tartozik a névhez; ez alapján a köznévi tagot tartalmazó neveket egy skálán tudjuk elképzelni.

A következő esetekben a név szerves része a köznévi utótag, nem hagyható el:

(19) *Váci utca*

(20) *Erzsébet híd*

(21) *Duna–Tisza köze*

A következő csoportban olyan nevek szerepelnek, melyeknél kérdéses, hogy a köznév hozzátartozik-e a tulajdonnévhez. Az ilyen nevek köznévi tagja a köznapi nyelvhasználatban gyakran elmarad, de mivel több lehetséges referens között egyértelműsít, információt vesztenénk, ha kihagynánk, ezért ezt is a névvel annotáljuk:

- (22) *Kent grófság*
- (23) *New York állam*
- (24) *Gyöngyös város(a)*
- (25) *Mátra hegység*
- (26) *Duna folyó*
- (27) *olasz Alpok*
- (28) *lengyel Magas-Tátra*
- (29) *Botond étterem*
- (30) *Keleti pályaudvar*

A képzeletbeli skála végén szerepelnek azok a kifejezések, amelyekben a tulajdonnév előtt alkalmi jelző áll:

- (31) *a gyönyörű Alpok*
- (32) „*Mit nekem te zordon Kárpátoknak...*”

Az alkalmi jelző nem része a tulajdonnévnek, itt tehát csak az *Alpok*, illetve a *Kárpátoknak* lesz tulajdonnévnek annotálva.

Annak, hogy meddig tart egy tulajdonnév, nemcsak az utána következő, nem szorosan hozzátartozó köznevek szabhatnak határt, hanem egy utána következő másik tulajdonnév is. Előfordulhat olyan eset, amikor egy más kategóriába tartozó név következik az egyik után, ilyenkor viszonylag egyszerű a különválasztás. De mi van olyankor, amikor ilyet látunk?

- (33) *Kovács János Bélával*

Ha minden kontextus nélkül látunk ehhez hasonló példát, akkor nem tudhatjuk, hogy a három nevet összevonhatjuk-e egy teljes személynévvé, vagy két, egy alanyesetű és egy instrumentális esetű névről van-e szó. Ilyen esetekben nagyon fontos a tulajdonnév ún. külső jegyeit is figyelembe venni,

mivel csak a belsők alapján ezt nem lehet eldönteni. McDonald (1996) a névfelismerés belső és külső bizonyítékait definiálja. Egy belső bizonyíték magából a nevet alkotó karakterláncból vezethető le, a külső bizonyítékok pedig a név kontextusából jönnek, abból a szövegbeli környezetből, amiben aktuálisan megtalálható.

5. A tulajdonnevek kompozicionalitása

Ebben a fejezetben annak a vizsgálatához, hogy a mono- és polimorfemikus tulajdonnevek kompozicionálisak vagy önkényesek-e, azt vizsgálom meg, hogy a megnevezett dologról való ismereteinket mennyire tudjuk a megnevezésből levezetni. A tulajdonnevek nem egyszerűen önkényes nyelvi jelek, hanem az önkényességet mint jelenséget szinte ezek mutatják a legvilágosabban: a kutyámnak vagy egy új használati tárgyamnak bármilyen nevet adhatok. Ebből a tényből, a névadás önkényességéből következik az is, hogy ezek a nevek semmit nem árulnak el a megnevezett dolog természetéről, sőt tulajdonképpen azt sem, hogy miről van szó, hiszen ugyanezeket a neveket bármi másnak is adhattam volna.

Bár a monomorfemikus tulajdonnevek a nem-kompozicionalitás iskolapéldái, azért szemantikailag ezek sem tökéletesen üresek. Például a *Charlie* név alapértelmezése fiú, bár amerikai nyelvterületen lánynak is gyakran adják, és természetesen adható háziállatnak, és más élőlénynek, sőt élettelen terméknek is. Vagyis a tulajdonnevek szemantikai implikációi, ha vannak is, felülírható (defeasible) jellegűek, éles ellentétben a köznevekkel, hiszen a sakkjátékot nem nevezhetjük malomnak, és a malmot nem nevezhetjük sakknak a kommunikáció grice-i minőségi maximájának megsértése nélkül. A monomorfemikus tulajdonneveknek csupán egy triviális nem-felülírható szemantikai implikációja van: ha valamit X-nek nevezünk, akkor arra igaz lesz az a predikátum, hogy a neve X.

A polimorfemikus tulajdonneveknek vagy tulajdonnévi csoportoknak két fajtáját különíthetjük el jelen vizsgálat szempontjából. Az egyik fajtába tartoznak azok a konstrukciók, ahol köznévi fejhez tulajdonnévi módosító kapcsolódik, mint például *Kossuth Lajos utca* vagy *Erzsébet-híd*. A másikba azok tartoznak, ahol mindkét közvetlen összetevő tulajdonnév: *Kossuth Lajos*, *Volvo S70*.

Az első (jóval gyakoribb) konstrukciótípus esetében minden szilárd (non-defeasible) szemantikai implikáció (kivéve az elnevezés tényét) a fejből következik, a módosító ehhez nem járul hozzá. Ez akkor válik igazán jól

láthatóvá, ha a fejet töröljük: „*a Bolyaiból hívnak*” mondatból nem derül ki, hogy akit a telefonhoz hívnak, azt a Bolyai Farkas Megyei Könyvtárból, a Bolyai János Gimnáziumból, a Bolyai utcai presszóból, vagy honnan keresik: csak a triviális implikáció marad meg, hogy annak a helynek Bolyai a neve. Hogy a konstrukció egészének szemantikájához a módosító mennyire nem járul hozzá, azt jól mutatja az is, hogy ebben a pozícióban teljesen üres elemeket (*A utca, B-híd*) is használhatunk anélkül, hogy a szerkezet egészének használhatósága bármiben is csorbulna. További érv a kompozicionalitás ellen az, hogy ha megpróbáljuk alkalmazni, elfogadhatatlan eredményekhez jutunk. A Széna téren nem árulnak szénát, a Boráros téren nem árulnak se borárosokat, se bort. A Kossuth Lajos utcában nem árulnak Kossuthot. A Váci út történetesen éppen Vácra vezet, de a Párizsi körút nem vezet Párizsba.

A második (tulajdonnévi fejet tartalmazó) konstrukció bonyolultabb: a magyarban egyébként szokatlan módon gyakran az előtag: a Volvo S70 egyfajta Volvo, és nem egyfajta S70. A konstrukció egyik legfontosabb példája a személynév, de itt sem mindig egyértelmű, hogy melyik a fej, és melyik a módosító: John Smithről gyanítjuk, hogy a Smith család a keresetségben John nevet kapott tagja, de Murazawa Takahashinál ezt már nem tudjuk biztosan, ahogyan Szulejmán ibn Abd al-Maliknál sem, hogy a keresetségről már ne is beszéljünk.

Általában tehát mindkét konstrukciónál csak annyit mondhatunk, hogy a jelentést az F fej hordozza, az M módosító szemantikai hozzájárulása csupán annyi, hogy a fej típusába tartozó egy bizonyos, M-nek nevezett F-ről van szó. Mindez éles ellentétben áll a köznévi módosítók megszokott kompozicionális szemantikájával, ahol a piros kalap olyan kalap, ami piros, a korábbi elnök korábban elnök volt, a hatalmas bolha (bolhának) hatalmas stb., és ezeket az implikációkat nem lehet felülírni.

Az annotálás gyakorlatára lefordítva ebből az következik, hogy mindig a leghosszabb nevet (a legkülsőbbet) jelöljük a jelölhetők közül. (Ebbe természetesen nem tartoznak bele a tulajdonnévhez kapcsolódó köznévi frázisok, tehát a *Kossuth Lajos utca bal oldalán valaha állt épület* frázisban csak a *Kossuth Lajos utcát* jelöljük tulajdonnévként.) Ebből következik, hogy nem annotálunk egymást átfedő vagy egymásba ágyazott neveket. Vagyis minden annotációnak be kell fejeződnie, mielőtt egy másik elkezdődik.

6. A tulajdonnevek toldalékolt alakjai

A toldalékolt alakok jelölését illetően több kódolási séma van forgalomban a tulajdonnév-felismerés terén dolgozók között. Mi azt a sémát követtük, amely szerint nem nyúlunk bele a morfológiai alak belsejébe, vagyis nem választható el a név a toldaléktól. Ebben az esetben vagy toldalékostul jelöljük az egész nevet, vagy egyáltalán nem jelöljük a toldalékolt alakokat.

Különbözőféleképpen kezeljük az inflexiós és a derivációs toldalékokat. Az inflexiós toldalékokról azt szokás gondolni, hogy nem változtatják meg alapvetően a szó alapjelentését, mint ahogy szófajt sem váltanak, ezért ha az azonosított tulajdonnév ragozott formában szerepel a szövegben, a toldalékkal együtt, a teljes alakot annotáljuk.

A képzőkkel más a helyzet: szófaji és ortográfiai változást okoznak, és jelentős mértékben megváltoztatják a név jelölését, sokszor egészen messze visznek az eredeti jelöltől. Ezért az ilyeneket nem annotáljuk tulajdonnévként:

(34) *fideszes*

(35) *Orbán Viktor-i*

(36) *gyurcsányozik*

(37) *petőfieskedő*

(38) *Top Gun-os*

(39) *Sass Tamás-féle*

A képzett alakok közül az egyetlen, amelyiket bevettük az annotálandók közé, az a helynév *-i/-beli* képzős alakja, és ezt is csak akkor jelöljük, ha a konkrét kontextusban helyre referál. A (40)-(41)-es példákban helynévként jelöljük a *budapesti*-t és a *romániai*-t:

(40) *a budapesti események*

(41) *a romániai Verespatakon levő bánya*

(A (41)-esben természetesen a *Verespatakon*-t is helynévként annotáljuk. Ez egy újabb példa arra az esetre, amikor több ugyanolyan kategóriájú egymás után következő nevet külön jelölünk.)

Az olyan mellékneveket, amelyek a jelölt dolognak nem a származására, hanem egyéb tulajdonságára, mondjuk elkészítési módjára vonatkozik, nem jelöljük tulajdonnévként. Például:

(42) *csípős szecsuáni mártás*

(43) *szegei halászlé*

Továbbá nem jelöljük a nemzetiségneveket sem, hiszen nem feltétlenül helyre referálnak, ahogy az *orosz hússaláta*, az *angol nyelv*, az *ukrán maffia* vagy a *magyar virscaft* példákból jól látható.

7. Metonimikus esetek

7.1. A metonímia definíciója

Akkor beszélünk metonímiáról, amikor egy kifejezést egy másik kifejezés helyett használunk bizonyos kontextusban. Például:

(44) *Az embereket sokkolta Vietnam.*

Ebben a mondatban egy földrajzi névvel, amely eredetileg egy helyre referál, jelen esetben egy eseményre utalunk, amely azon a helyen történt. Hasonlóan a további példákban:

(45) *Az Eötvös József Gimnázium nem kap elegendő állami támogatást.*

(46) *Nincs messze tőlünk az Eötvös József Gimnázium.*

(47) *Az Eötvös József Gimnázium idén Luxemburgba megy kirándulni.*

A (45)-ös példamondatban az intézménynév ténylegesen egy intézményre utal, míg a (46)-osban már konkrét fizikai helyre, az (47)-esben pedig egy emberi közösségre referálunk ugyanazzal a névvel.

A metonímiákban tehát egy fogalmat vagy dolgot egy másik fogalom vagy dolog jelölésére használunk (Kövecses 2005). Referenciaátvitel történik: egy névvel az eredeti referens helyett egy másik referensre utalunk. A hagyományos nyelvtanok a metonímiát jelentésátvitelként definiálják – én szándékosan tartózkodom ettől a kifejezéstől. A tulajdonnevek esetében a jelölő—jelölt—jelentés hármassal legalábbis nem problémamentes. A különféle nyelvészeti irányzatok követői különféleképpen nyilatkoznak a tulajdonnevek jelentéséről; jó példák találhatók erre (J. Soltész 1979)-ben, (Kiefer 2000)-

ben és (Antal 1978)-ban. Én azt az irányzatot követem, amely azt mondja, hogy a tulajdonneveknek nincs jelentésük, csak jelöletük, más szóval denotátumuk. Ennek megfelelően jelentésátvitelről sem beszélhetünk, ezért használok a referenciaátvitel kifejezést.

A metonímiák feloldásának fontosságát a természetesnyelv-feldolgozás több különböző területén is kimutatták, így a gépi fordításban (Kamei és Wakao 1992), a kérdésmegválaszoló rendszerekben (Stallard 1993), az anaforafeloldásban (Harabagiu 1998; Markert és Hahn 2002) és persze a tulajdonnév-felismerésben (Markert és Nissim 2007; Farkas R. és mtsai 2007).

7.2. *A metonímiák csoportosítása*

Bár a potenciális metonimikus olvasatok száma végtelen, és a metonimikus nyelvhasználat nagyon újító és termékeny, bizonyos minták azért kirajzolódnak. Az eddig bemutatott példák esetében szabályos poliszémiáról beszélhetünk, mert minden esetben van legalább még egy név, aminek a referenciái ugyanígy különböznek. Sőt a poliszémia általában egy-egy szemantikai mezőbe tartozó kifejezésekre vonatkozik. Markert és Nissim (2006) az utóbbi csoportot, amelyek esetében a metonimikus esetek konkrét mintázatok alapján szerveződnek, *konvencionális* metonímiáknak, míg a sémákba nem rendezhető egyedi darabokat *újszerű, nem konvencionális* metonímiáknak nevezik.

A metonímiák nemzetközileg elfogadott jelölési módja az A-FOR-B formula, ahol az A kifejezés áll a B kifejezés helyett. Például a PLACE-FOR-PEOPLE metonímiában helynévvel referálunk egy emberi közösségre. A továbbiakban Kövecses (2005) magyar formuláját fogom használni, aminek a sémája: AZ X AZ Y HELYETT, vagyis a fenti angol példa magyarul: A HELY AZ EMBEREK HELYETT.

7.2.1. *Az osztályfüggetlen olvasatok*

Az osztályfüggetlen olvasatok közé azokat a metonimikus mintázatok sorolja Markert és Nissim (2006), amelyek minden típusú tulajdonnévre alkalmazhatóak (és minden konkrét főnévre is, de itt most csak a tulajdonnevekről lesz szó). Az egyik ilyen minta az A TÁRGY A NÉV HELYETT (OBJECT-FOR-NAME). Ebben az esetben tulajdonképpen a név metanyelvi használatáról van szó, amikor a konkrét névalakról beszélünk, például:

(48) *Nekem tetszik a Dávid név.*

AZ A TÁRGY A REPRESENTÁCIÓ HELYETT (OBJECT-FOR-REPRESENTATION) metonímia esetében a névvel annak valamilyen reprezentációjára utalunk:

(49) *Ez itt Shakespeare.* (egy Shakespeare-ről készült képre mutatva)

(50) *Málta itt van.* (a térképre mutatva)

Minden típusú névvel és szimpla főnevekkel is előállhat az az eset, amikor a mondatban betöltött különböző grammatikai szerepeik ütközése miatt vegyes olvasatot kapunk:

(51) *A három balti ország – Észtország, Lettország, Litvánia – részvételével tegnap megkezdődött a konferencia.*

Az (51)-esben az országnevek egyszerre literális olvasatú helynevek (mert ott van az *a három balti ország* frázis) és A HELY AZ EMBEREK HELYETT metonímiák (mivel a konferenciákon emberek szoktak ülni).

7.2.2. *Az osztályspecifikus olvasatok*

A HELY VALAMI HELYETT

A helynevek közé a geográfiai és/vagy politikai-közigazgatási alapon definiált földrajzi egységek nevei tartoznak (országok, városok, megyék stb.), melyek jelölhetnek egy helyet, egy kormányzatot, egy közösséget vagy akár az adott terület iparát is. Minden olyan esetben, amelyben a helynév aktív cselekvői pozíciót tölt be a mondatban, vagyis fizikai mozgást végez vagy indít, döntést hoz vagy érzelmei vannak, A HELY AZ EMBEREK HELYETT (PLACE-FOR-PEOPLE) metonímiáról beszélünk.

(52) *Franciaország korlátozza a politikai menedékjogot.*

(53) *Franciaország új elnököt választott.*

Ennek a metonímiának egy nagyon tipikus és sűrűn használt alosete az, amikor a hely nevével egy sportcsapatra utalunk. Például:

(54) *A Manchester ma a Münchennel játszik.*

(55) *Olaszország nyerte a foci végét.*

A metonímia definíciója kapcsán bemutatott példában helynév szerepel egy esemény helyett; ilyen akkor fordul elő, amikor egy esemény nagyon erős asszociációs viszonyban áll egy adott hellyel. Ezeket hívjuk A HELY AZ ESEMÉNY HELYETT (PLACE-FOR-EVENT) metonímiáknak, például:

(56) *Trianon megítélése a két háború közötti időben*

(57) *Federer idén is meghódította Wimbledon.*

A SZERVEZET VALAMI HELYETT

A helynevekhez hasonlóan a szervezetnevek is több dologra tudnak referálni az eredeti referens mellett. A leggyakoribb típus az A SZERVEZET A TAGOK HELYETT (ORGANISATION-FOR-MEMBERS) metonímia, ami olyan esetekben szokott előállni, amikor a szervezetnév a mondat aktora, kommunikációs aktusokat tesz, emocionális, illetve mentális állapotai vannak, döntéseket hoz, tervei, céljai vannak. Mivel ilyeneket jellemzően csak emberek csinálnak, ezért minden ilyen esetet metonímiaként kell számon tartanunk.

(58) *Az IBM ma jelentette be új technológiáját.*

(59) *Az apcmag.com egyik cikke szerint a Microsoft elnézést kért.*

A szervezeteknek nemcsak felépítési struktúrájuk van, hanem székhelyük is, ezért sokszor előfordul a szövegben, hogy a helyre a szervezetnévvel utalunk. Ilyet tapasztalunk például a cégneveknél vagy a kormányzati hivataloknál, amelyeknek jellemzően egy épületben van a székhelyük. Ilyenkor a szervezetnévvel utalunk az épületre, vagyis ez A SZERVEZET AZ ÉPÜLET HELYETT (ORGANISATION-FOR-FACILITY) metonímia.

(60) *A János kórházban sok a macska.*

(61) *A Nemzeti Múzeum az 1848. március 15-ei események egyik fő helyszíne volt.*

További meglehetősen gyakori metonímiatípus az A SZERVEZET A TERMÉK HELYETT (ORGANISATION-FOR-PRODUCT), amikor jellemzően egy cég által gyártott termékre a cég nevével utalunk. Ezek a lehető leghétközna-

píbb példák, amikor már tényleg nem vesszük észre, hogy bármiféle referenciaátvitel történt:

(62) *Egy Volvo kormánya mögött érezhető igazán a kényelem és a dinamika.*

(63) *Kairó utcáin még mindig sok a hatvanas évekből származó Renault.*

Szintén meglehetősen jellemző, főleg gazdasági rövidhírekben és tőzsdei jelentésekben gyakran előforduló metonímia az A SZERVEZET AZ INDEX HELYETT (ORGANISATION-FOR-INDEX):

(64) *A Mol 10 forinttal 6640 forintra, míg a Matáv 1 forinttal 823 forintra csúszott vissza.*

7.3. A metonímiák annotálása

A konvencionális metonímiák egy-egy teljes szemantikai mezőre vonatkoznak, jellemzőek és megjósolhatóak, ezért valamilyen konzekvens jelölési módot kell rájuk kitalálni egy tulajdonnév-felismerő alkalmazás céljára épülő korpusz annotációs rendszerében. Két elv ismert és használt a tulajdonnév-felismerés területén a metonimikus esetek kezelésére.

Ha a *tag-for-meaning* elvét alkalmazzuk, akkor a kontextusnak megfelelően, az éppen aktuális referens címkéjét kapja a név, például:

(65) *A Szépművészeti Múzeumban elszaporodtak a patkányok.*

(66) *Új kiállítást nyit a Szépművészeti Múzeum.*

Ezt az elvet követve a (65)-ösben helynévi, a (66)-osban szervezetnévi címkét kapna ugyanaz a név.

Egy másik elv, a *tag-for-tagging* elve alapján viszont egy név kontextustól függetlenül mindig ugyanazt a címkét kapja, vagyis a kiinduló referensét. Ebben az esetben mindkét fenti példamondatban szereplő név ugyanúgy a szervezetnévi címkét kapja. Ahhoz, hogy ezt az elvet értelmesen tartani lehessen az annotálás során, elképzelhető, hogy új névosztályokra lesz szükség. Egy új kategóriarendszer létrehozásával próbálkoztak az Automatic Content Extraction (ACE) konferencia annotálási sémájának kidolgozói (ACE 2004, 2005, 2007), akik azt a problémát, hogy egy helynév utalhat földrajzi egységre, emberekre, eseményre, úgy oldották meg, hogy bevezették a geográfiai/politikai/szociális entitások kategóriáját, amibe például egy országnév minden használati módjában belefér.

A HunNER korpuszban egy – a fenti elvek mindegyikének eleget tevő – harmadik megoldást alkalmazunk: jelöljük az eredeti referens típusát és a metonímia tényét és típusát is.

8. Összefoglalás

A tulajdonnevek definiálása és jelentésük meghatározása a nyelvészetben komoly problémákat okoz. Sokan sokféleképpen próbálták megragadni a tulajdonnév fogalmát, de ezek a próbálkozások nagyrészt az elmélet szintjén maradtak, holott egy olyan gyakorlati területen, mint a számítógépes tulajdonnév-felismerés is sok nehézséget okoznak. Bármilyen szövegből is akarunk információt kinyerni a számítógép segítségével, mindenhol találkozunk tulajdonnevekkel, melyek sok szempontból hasonlóan viselkednek a sima főnevekhez, sok tulajdonságukban viszont eltérnek tőlük. Dolgozatomban egy tulajdonnév-felismerő rendszer és az ahhoz szükséges korpusz építése során előforduló nyelvészeti problémákat igyekeztem áttekinteni – a teljesség igénye nélkül, hiszen a körüljárt témák mellett számos egyéb érdekes kérdés is felmerül a nevekkel kapcsolatban, melyek tárgyalása más dolgozatok témája lehet.

A dolgozat legnagyobb fejezete a nevek metonimikus viselkedésével foglalkozik, mely az elmúlt évtizedekben a kognitív metaforaelméletek megerősödésével egyre inkább előtérbe került a nyelvészeti kutatásokban. A nem szó szerinti használat természetesen nem csak a nevek sajátja, de míg a köznevek kapcsán ezen a területen elég nagyszámú kutatás folyik, a tulajdonnevekhez kevesen nyúltak hozzá. Korpuszépítési munkánkat folytatva tovább dolgozunk a HunNER korpusz fejlesztésén, és emellett belekezdünk egy olyan magyar nyelvű korpusz építésébe, melyben automatikus eszközökkel fogjuk bejelölni a metaforikus kifejezéseket és neveket, további információkat gyűjtve ezzel a tulajdonnevek tulajdonságairól.

HIVATKOZÁSOK

The ACE {2004, 2005, 2007} Evaluation Plan.

Elérhető: <http://www.nist.gov/speech/tests/ace>

Antal L. 1978: *A jelentés világa*, Magvető Kiadó, Budapest.

- Chinchor, N. – Robinson, P. 1998: MUC-7 Named Entity Task Definition Version 3.5, in *Proceedings of the 7th Message Understanding Conference (MUC-7)*. Elérhető: http://www-nlpir.nist.gov/related_projects/muc/proceedings/ne_task.html
- Farkas R. – Simon E. – Szarvas Gy. – Varga D. 2007: GYDER: maxent metonymy resolution, in *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, Prague.
- Harabagiu, S. 1998: Deriving metonymic coercions from WordNet, in *Workshop on the Usage of WordNet in Natural Language Processing Systems, COLING ACL*, 142–148.
- J. Soltész K. 1979: *A tulajdonnév funkciója és jelentése*, Akadémiai Kiadó, Budapest.
- Kamei, S. – Wakao, T. 1992: Metonymy: Reassessment, survey of acceptability and its treatment in machine translation systems, in *Proceedings of ACL*, 309–311.
- Kiefer F. 2000: A szöösszetétel, in Kiefer F. szerk. *Strukturális magyar nyelvtan 3: Morfológia*, Akadémiai Kiadó, Budapest.
- Kiefer F. 2000: *Jelentéselmélet*, Corvina, Budapest.
- Kövecses Z. 2005: *A metafora. Gyakorlati bevezetés a kognitív metaforaelméletbe*, Typotex, Budapest.
- Markert, K. – Hahn, U. 2002: Understanding metonymies in discourse, *Artificial Intelligence* 135, 145–198.
- Markert, K. – Nissim, M. 2006: Metonymic Proper Names: A Corpus-based Account, in Stefanowitsch, A. – Gries, Th. ed.: *Corpus-based approaches to metaphor and metonymy*, Mouton de Gruyter.
- Markert, K. – Nissim, M. 2007: SemEval-2007 Task 08: Metonymy Resolution at SemEval-2007, in *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, Prague.
- McDonald, D. 1996: Internal and external evidence in the identification and semantic categorization of proper names, in Boguraev, B. – Pustejovsky, J. ed.: *Corpus Processing for Lexical Aquisition 2.*, MIT Press, Cambridge, MA, 21–39.
- Simon E. – Farkas R. – Halácsy P. – Sass B. – Szarvas Gy. – Varga D. 2006: A HunNER korpusz, in Alexin Z. – Csendes D. szerk.: *IV. Magyar Számítógépes Nyelvészeti Konferencia*, Szeged.
- Stallard, D. 1993: Two kinds of metonymy, in *Proceedings of ACL*, 87–94.